Editor's note: This is the 113th in a series of review and tutorial papers on various aspects of acoustics.

On automated language acquisition

Allen Gorin[a)]
AT&T Bell Laboratories, Murray Hill, New Jersey 07974

(Received 18 June 1994; revised 13 December 1994; accepted 22 December 1994)

The purpose of this paper is to review our investigation into devices which automatically acquire spoken language. The principles and mechanisms underlying this research are described and then experimental evaluations for several tasks are reported, involving both spoken and keyboard input. The generic mechanism in these experiments is an information-theoretic connectionist network embedded in a feedback control system.

PACS numbers: 43.10.Ln, 43.72.Ne

INTRODUCTION

We are interested in devices which understand and act upon spoken input from people. Traditionally, in such speech understanding systems, the hierarchy of linguistic symbols and structures has been manually constructed, involving much labor and leading to fragile systems which are not robust in real environments. In human language acquisition, however, the phonemes, vocabulary, grammar, and semantics seem to emerge naturally during the course of interacting with the world. This contrast motivates us to investigate devices which automatically acquire the language for their task, during the course of interacting with a complex environment. While a long-term investigation, research in such *language acquisition* devices yields insights into how to construct speech understanding systems which are trainable, adaptive, and robust. The purpose of this paper is to recount our progress and ideas to date in this endeavor. In particular, we describe the principles and mechanisms underlying this research and review several experimental systems which have been constructed.

A *first principle* in our research is that the purpose of language is to convey meaning, so that language acquisition crucially involves learning to decode that meaning. A *second principle* is that language is acquired during interaction with a complex environment, wherein the device receives some input stimuli, responds to that input, then receives feedback as to the appropriateness of its response. These principles underlie our investigation into *connectionist* mechanisms, in which a network constructs associations between input stimuli and appropriate machine responses. We embed these networks in a *control-theoretic* mechanism for governing language acquisition via *reinforcement learning*. If the reinforcement feedback is positive, then the associations are strengthened, while negative reinforcement causes the associations to be weakened.

A system block diagram based on these principles is shown in Fig. 1. The device receives some input, comprising linguistic and possibly other stimuli. In response to this input, it performs some action, to which its environment then provides a *semantic-level* error signal as to the appropriateness of that response. The device then *adapts* its behavior based on this error feedback. Thus we assume that the system will make occasional errors, especially when encountering unfamiliar stimuli. The emphasis, however, is on its ability to *detect* an error via reinforcement feedback from the environment, to *recover* from the error via feedback control, then finally to *learn* from the error so that it is not repeated.

The goal of man–machine communication, in such a system, is to induce the machine to undergo some transformation. This transformation can be immediately observable in the form of some machine action, or can be an internal state change which is only observable indirectly based on some future interaction. We denote the input to the device as *language* and the mapping from input to transformation as *understanding*. We are then satisfied that the machine understands if it responds appropriately over a wide range of input scenarios, which is essentially a reformulation of the Turing Test.

It is worthwhile contrasting this paradigm with traditional communication theory, best accomplished by a quotation from Shannon's original paper (emphasis added) (Shannon, 1948):

> *"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently these messages have* meaning; *that is they refer to or are correlated accordingly to some system with certain physical or conceptual entities. These semantic aspects of communication are* irrelevant *to the engineering problem."*

In contrast, for systems which purport to understand spoken language, the *semantic* aspects of communication are primary. How then can we quantify such notions? In people, an input stimulus evokes memories of associated perceptions and activities. We thus propose a *third principle*, that meaning is grounded in a device's interaction with its environment. This principle underlies an investigation of methods to quantify the meaning of spoken language via its network associations to a device's input/output periphery, providing an acquired representation of the device's operational environment. Introducing a metric and norm on these associations form the basis of a *salience* theory, which quantifies the

[a)] E-mail:algor@research.att.com

0001-4966/95/97(6)/3441/21/$6.00

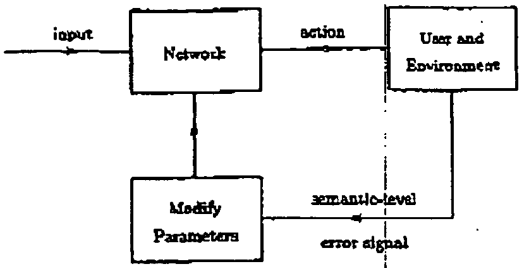

FIG. 1. Reinforcement learning in language acquistion.

information content of an input stimulus for a particular device.

In the remainder of this Introduction, we motivate and outline the algorithmic methods which have proved useful in our investigation of such language acquisition devices. The body of the paper will describe in detail these mechanisms and experimental evaluations thereof.

An information-theoretic network. For the system illustrated in Fig. 1, the issue arises of how to implement the mapping from input stimuli to action. We have proposed constructing *connectionist* networks which build associations between input stimuli and appropriate machine responses (Gorin *et al.*, 1991). If a machine receives positive reinforcement, then the connections are strengthened, while negative reinforcement causes the connections to be weakened. There are many methods that have been proposed in the literature for learning such connection weights. In particular, we define the connection weights of these networks via *mutual information*, which has a variety of theoretical and practical advantages over gradient-based training methods. The definition and properties of such *information-theoretic networks* will be described in Sec. I.

Our earliest experiments involved a single-layer information-theoretic network which constructs direct associations between words and meaningful machine actions. This simple architecture corresponds to a "bag-of-words" language model. It was then extended to a multi-layer network which in addition builds associations between phrases and actions, acquiring a rudimentary syntactic structure to improve understanding.

Those early experiments involved a text-based Automated Call Routing system (Gorin *et al.*, 1991), then a spoken-input version of that same system (Gorin *et al.*, 1994a). That scenario involved a Department Store, which receives input such as *I need some paint for my redwood table*, whence the appropriate response is to route the caller to the Hardware Department. More recently, these methods were applied to a database of actual customer/operator dialogs from the AT&T telephone network (Gorin *et al.*, 1993b) (Gorin *et al.*, 1994b) (Sankar *et al.*, 1993). In this scenario, an input might be *I want to reverse the charges*, whence the appropriate response is to route the caller to an automated subsystem which handles collect calls. These experimental systems will be described in Sec. III.

Structured networks. As a device and its task become more complex, so does the mapping from input stimuli to machine action. Given some network architecture, a reasonable question is to ask whether it is capable of learning such complex mappings? A striking feature of human language acquisition is our ability to make sweeping generalizations from small numbers of observations. For example, a single observation of a new word, in the appropriate context, can suffice to acquire its pronunciation, syntactic role, and semantic associations.

We observe that the neural network in a biological organism is *not* homogeneous, but rather highly structured and modular. Such structure develops over evolutionary time, matching itself to a species' sensory/motor periphery and environment. One can hypothesize that the constraints provided by such network structure correspond to the innate characteristics which enable an individual organism to rapidly adapt to its environment (Bunge, 1986). This motivates a *fourth principle*, that in order to provide rapid learning and generalization a device must reflect the structure of its input/output periphery and environment. As observed in (Minsky and Papert, 1990),

> *"the marvelous powers of the brain emerge not from any single, uniformly structured connectionist network but from highly evolved arrangements of smaller, specialized networks which are interconnected in very specific ways."*

Thus motivated, we have investigated *structured* network architectures whose constraints greatly accelerate the learning process. In particular, we developed several methods for constructing large structured networks by combining component subnetworks, then experimentally evaluated those networks in several application scenarios.

In a Call Routing task, the set of machine actions is merely a list, corresponding to a particularly simple output periphery structure. One can consider the more general situation where the machine actions comprise an *n* parameter set of subroutine calls. Miller has proposed the construction of *product networks* for such devices, where individual subnetworks are allocated to each output parameter, thereby reflecting the output periphery structure in its network architecture (Miller and Gorin, 1993c). This product network enables improved generalization by factoring phrase/action associations through intermediate *semantic primitives*. These ideas will be expanded and detailed in Sec. IV.

A two-dimensional product network has been experimentally evaluated on an *Almanac* data retrieval task, first text-based (Miller and Gorin, 1993c), then speech-based (Gorin *et al.*, 1993a) (Miller and Gorin, 1993b). The system responds to inputs such as *What is the largest mountain in the Empire State?*, to which the appropriate machine response is *The highest point in New York State is Mt. Marcy (5,344 feet)*. This experimental system will be described in Sec. IV.

In many situations of interest, the appropriate machine response to a spoken input depends not only on the message, but also on the state of its environment. This motivates us to investigate devices with both linguistic and other input channels. Such extra-linguistic information can serve to resolve ambiguities during understanding as well as provide redundancy to accelerate language acquisition. For such devices, Sankar has proposed the construction of *sensory primitive subnetworks*, which learn the cross-channel associations between different input stimuli (Sankar and Gorin, 1993). These subnetworks are combined in a product architecture, reflecting a device's input periphery structure, whose output is then used to control the machine actions. This architecture provides improved generalization by factoring phrase/action associations through the sensory primitive subnetworks.

Sankar evaluated this network and control strategy in a Blocks World scenario, in which the machine has both linguistic and visual input channels (Sankar and Gorin, 1993). It is presented with a scene comprising objects of varying color and shape. The machine actions comprise *focusing its attention* on a particular object in response to input such as *Where is the red square?* Such focus of attention is a necessary prerequisite to more complex actions. More recently, Henis has extended this system, connecting it to a robotic simulator where the machine actions comprise manipulating the blocks upon which it has focused its attention (Henis *et al.*, 1994). These experimental systems will be discussed in Sec. IV.

Symbols from signals. In order to provide rapid learning and generalization in a language acquisition device, we have explored methods for reflecting the structure of a device's input/output periphery in its network architecture. There is also structure in the environment, in that the input signals to a device can be organized into symbols, then further organized into hierarchical structures. While such structure can be imposed on all sensory inputs, in this research we focus on linguistic symbols and structures. We believe, however, that our methods are general and can be applied to all cognitive modalities.

We focus initially on words, which are the fundamental symbols of meaning in language. In traditional speech recognition systems, one specifies the vocabulary *a priori* then trains the recognizer by presenting it with labeled speech (Rabiner and Juang, 1993). During human language acquisition, however, words seem to emerge naturally during the course of interacting with the world. How might this be? Furthermore, how might we mimic such characteristics in our devices so as to improve their trainability, adaptability, and robustness?

This contrast motivates us to investigate methods for automated acquisition of spoken words. Webster (Webster, 1987) defines a word as

> *"a speech sound ... that communicates meaning ... without being divisible into smaller units capable of independent use."*

Based on the intuition that a symbol should be a stable point of some operator, we have investigated clustering algorithms that search for speech sounds which are acoustically and semantically consistent. A prerequisite to measuring such consistency is to define acoustic and semantic feature spaces with appropriate metrics.

In people, an input stimulus evokes memories of associated perceptions and activities. This motivated us to define the *meaning* of a word, for a particular device, to be its *network associations* to the device's input/output periphery. Such a definition grounds meaning in a device's interaction with its world, being dependent on its input/output periphery, environment and experiences. In Sec. V we will describe illustrative examples of such *semantic/sensory associations* for several experimental devices.

We have defined a distance between these association vectors, measuring the semantic similarity of two words for a device. We furthermore define a norm, which measures the semantic significance of an individual word or phrase. This norm measures the *information content of a word for the device*, which we denote *salience*. This can be distinguished from and compared to the traditional Shannon measure of *information content*, which measures the uncertainty that a word will occur. These theoretical and empirical relationships will be discussed in Sec. V.

Based on these ideas, we constructed and evaluated a rudimentary spoken language understanding system (Gorin *et al.*, 1994a). It is unique in that *no* text is provided to the device during either testing or training, in contrast to all other speech understanding systems. It is also unique in that the vocabulary and grammar are *unconstrained*, being acquired by the device during the course of performing its task. This is also in contrast to all other systems, where the salient vocabulary words and their meanings are explicitly provided to the machine. The initial application vehicle for this experiment in spoken language acquisition was the Department Store task (Gorin *et al.*, 1994a), then the Almanac (Gorin *et al.*, 1993a) (Miller & Gorin, 1993b). These experimental systems will be described in Sec. VI.

Grammatical inference. The above experiments focused on acquiring word symbols from the speech signal. The next level up in the linguistic hierarchy is grammar, comprising symbols and structure which govern the acceptable combinations of words into sentences. Grammar plays two important roles in speech understanding. First, it constrains the allowable word sequences, increasing the signal-to-noise ratio and thus improving our ability to recognize words in noisy or highly variable environments. Second, it modulates the meaning of a word according to its position in a sentence. Thus meaning can be viewed as an attribute of a word in a particular syntactic state, rather than of the word alone.

The automated acquisition of grammar has received much attention, intertwined with the classical debate concerning how much of linguistic structure must be innate in order to account for human behavior. We observe, however, that the acquisition of grammar from merely listening to speech is a much harder problem than people actually solve. In humans, language is acquired during the course of interacting with the world, exploiting both speech and other sensory input. A challenge, then, is to understand how to exploit such extra-linguistic information to guide grammatical inference, governed by the goal of learning to decode meaning.

As motivation, let us consider the basic parts-of-speech such as nouns and verbs. In elementary school, children are taught that a noun is a *person, place, or thing*, and that a verb is a word that expresses *an action*. It is striking that the classroom definition of such fundamental syntactic concepts are purely semantic. If one constructs a machine that can interact with things, for example in a Blocks World, then all phrases with high salience for such things can be clustered into a part-of-speech. Such an abstraction would correspond to the early semantic characterization of a noun. Similarly, given a machine which can sense the attributes of things (e.g., color or shape), then one could acquire a part-of-speech corresponding to the early notion of an adjective. Verbs could similarly be emergent from associations to time derivatives of such attributes.

While such definitions are a subject of much debate in linguistics, they serve as useful intuitions to motivate our investigation into exploiting semantic/sensory associations for grammatical inference. In Sec. VII, we will first describe the method of *salience thresholding* in an information-theoretic connectionist network. This thresholding yields a subnetwork which corresponds to a part-of-speech for each dimension of the device periphery. The resultant subnetwork is activated only by those words or phrases which are highly salient for its semantic or sensory primitive.

Once these parts-of-speech are acquired, they can be manipulated just like any other symbol. We thus propose a *fifth principle*, that language acquisition proceeds in developmental stages, from the concrete to the abstract, from the simple to complex.[1] In adult language, parts-of-speech are characterized both via their meaning and within-language usage patterns. In Sec. VII, we also report on preliminary experiments which re-estimate induced parts-of-speech so that they become consistent from both these perspectives.

An application of these ideas was explored by Gertner, who constructed a hierarchical network with subnetworks corresponding to parts-of-speech in an Airline Information task (Gertner and Gorin, 1993). A query to that system might be *I want to leave New York and fly to the Windy City*, to which the appropriate machine response would be to display a flight table from New York to Chicago. The principle of developmental learning tells us that in order for a device to acquire the language involving pairs of places, it must first acquire the language associated with individual places.

A stable subnetwork for places was embedded in a hierarchical network, with secondary subnetworks corresponding to modifier phrases. Rapid learning and generalization was achieved by factoring phrase/action associations through these modifier subnetworks. For example, an encounter with the phrase *leave New York* leads to the acquisition of the meaning of *leave* as it relates to all place names.

Summary. The principles and mechanisms presented here form the basis of a theory of syntax and semantics, where conveying meaning is primary and linguistic structure serves to make such communication robust. Although our experimental devices are thus far rudimentary, we consider them to be the early stages of a long-term investigation into machines which automatically acquire language through interaction with a complex environment.

This paper proceeds as follows. Section I defines the basic information-theoretic network, its training procedure and basic properties. The feedback control mechanism used for dialog control is described in Sec. II. In Sec. III, we describe the experimental evaluation of that basic network in Call Routing tasks. Structured networks are discussed in Sec. IV, in particular their application to the Almanac and Blocks World tasks. In Sec. V, we define salience, discussing its relationship to information theory and providing several illustrative examples of semantic/sensory association vectors. Our experiments in spoken language acquisition are summarized in Sec. VI, where the device acquires spoken words from speech with no intervening text. In Sec. VII, rudimentary experiments in salience-based grammatical inference are discussed. The application of these ideas to an Airline Information task will then be described, involving a structured hierarchical information-theoretic network.

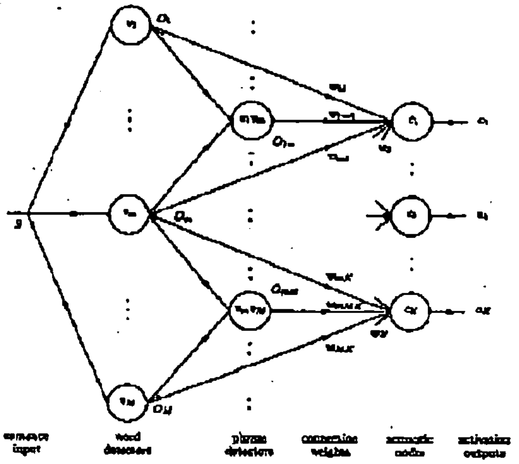

FIG. 2. A multilayer network mapping language to semantic actions.

I. AN INFORMATION-THEORETIC CONNECTIONIST NETWORK

In this section, we describe a mechanism for learning the mapping from input stimuli to machine action. In particular, we describe a connectionist network, originally proposed in (Gorin *et al.*, 1991), which builds associations between input stimuli and appropriate machine responses. If the machine receives positive reinforcement to a response then the connections are strengthened, while negative reinforcement causes the connections to be weakened.

The basic network architecture is illustrated in Fig. 2. A spoken or typed sentence is applied to the input layer, which comprises a collection of word-detector nodes. These nodes produce an output between zero and one, approximating the probability that a particular word is present in the input. In the simplest case for keyboard input, the output equals one if the input word exactly matches the node, else zero.

The intermediate layer comprises a collection of phrase-detector nodes, in the simplest case corresponding to adjacent word pairs (bigrams). The output layer comprises nodes which correspond to the various actions that the machine can perform. In this discussion, the action space is a list of subroutine calls. The extension of these methods to more complex devices via structured networks is discussed in Sec. IV.

The input and intermediate layers are shown partially grown. They are initialized at zero, growing over time as new words and phrases are encountered. In the case of keyboard input, a word token can be defined as any character sequence delimited by blanks or punctuation. A *new* word can be most simply defined as one which differs in any way from the existing vocabulary (Gorin *et al.*, 1991). This new word criterion can be softened (Miller and Gorin, 1993c) based on string-distortion measures such as the Levenshtein distance (Levenshtein, 1966) (Sankoff and Kruskal, 1983). In Sec. VI, we address the related issues for spoken input (Gorin *et al.*, 1994a), which involve both acoustic and semantic distortions.

There have been a number of methods proposed in the literature for training such networks (Rumelhart and McClelland, 1988). In this research, we have defined the connection weights between words and actions to be the mutual information between those events (Cover and Thomas, 1991) This leads to several attractive properties, as discussed later in this section. If we denote the current vocabulary of N words by $V=\{v_1,v_2,...,v_N\}$ and the set of K actions by $C=\{c_1,c_2,...,c_K\}$, then the information-theoretic connection weights are given by

$$w_{nk}=I(v_n,c_k)=\log_2\frac{P(c_k|v_n)}{P(c_k)}, \qquad (1)$$

where $P(c_k|v_n)$ is the conditional probability that a sentence containing word v_n connotes action c_k, where $P(c_k)$ is the prior probability of that action, and $I(v_n,c_k)$ is the mutual information between the word and action. This is intuitively satisfying as follows. If the presence of word v in a sentence makes an action c more likely, then $P(c|v)>P(c)$, so that the connection weight is positive (excitatory). Similarly, if the word v makes an action c less likely, then the connection weight is negative (inhibitory). Finally, if the word has no effect, then the conditional and prior probabilities are equal, so that the connection weight is zero (null).

The connection weight between a phrase and action is defined via excess mutual information. While in principle scalable to any n-gram phrase or set thereof, we restrict this discussion to adjacent word pairs (v_i,v_j).

$$w_{ijk}=I(v_iv_j,c_k)-I(v_i,c_k)-I(v_j,c_k). \qquad (2)$$

Biases for each output node (cf. Fig. 2) are given by

$$w_k=\log_2P(c_k). \qquad (3)$$

The activation at each output node is computed via a linear combination of those inputs,

$$a_k=\sum_{ij}\mathcal{O}_{ij}w_{ijk}+\sum_n\mathcal{O}_nw_{nk}+w_k, \qquad (4)$$

where $\mathcal{O}_{ij}$ is the output from the phrase-detector node for v_iv_j and $\mathcal{O}_n$ is the output of the word-detector node for v_n. That action c_{k_1} which has maximum activation is then performed, where

$$k_1=\arg\max_k a_k. \qquad (5)$$

Theoretical properties. The information-theoretic network has several relationships to other methods, which we briefly review here. Given suitable Markovian assumptions on the language, then the above algorithm is equivalent to a *maximum a posteriori* (MAP) decision (Gorin *et al.*, 1991). Given suitable independence assumptions, then a bag-of-words model applies and the connections from the intermediate layer vanish, yielding a single layer network. Under those conditions, the network is again equivalent to a MAP decision (Gorin *et al.*, 1991). Tishby observes that the information-theoretic network is equivalent to the criterion of classification via minimum description length where that interpretation is selected which provides for the minimum code length of the input sentence (Tishby and Gorin, 1994)

Addressing the problem of rule-inference for expert systems, Goodman shows that the strength of a candidate rule can be characterized by the mutual information between the *if* and *then* clauses (Goodman *et al.*, 1992). He furthermore describes a method for combining the parallel firings of such rules via a connectionist network with information-theoretic weights. This result is rather satisfying, ameliorating the traditional debate between connectionist and rule-based approaches to machine intelligence.

Tishby proves a universality theorem for information-theoretic associations, showing that, under suitable hypotheses, any association measure which is functionally related to probabilities can be rescaled to mutual information (Tishby and Gorin, 1994). There is a seemingly related set of results proving that when appropriately trained via a mean-squared error (MSE) criterion, the network outputs provide estimates of *a posteriori* probabilities (Richard and Lippmann, 1991). We remark that, while all these relationships are quite fascinating, fully understanding and exploiting them remains an issue for future research.

Estimation. After having decided on the information-theoretic network, the issue remains of how to estimate mutual information. The probabilities in formulas (1) through (3) can be estimated via smoothed relative frequencies (Gorin *et al.*, 1991). In particular, after encountering input sentences $s_1,s_2,...,s_\ell$, let $N_\ell(c_k,v_n)$ denote the number of sentences of class c_k containing word v_n, and let $N_\ell(c_k)$ denote the number of sentences in that class. We compute (Gorin *et al.*, 1991) smoothed relative frequency estimates of $P(c_k)$ and $P(c_k|v_n)$ via

$$\hat{P}_\ell(c_k)=(1-\alpha_\ell)\frac{1}{K}+\alpha_\ell\frac{N_\ell(c_k)}{\ell}. \qquad (6)$$

$$\hat{P}_\ell(c_k|v_n)=(1-\beta_\ell)P(c_k)+\beta_\ell\frac{N_\ell(c_k,v_n)}{N_\ell(v_n)}. \qquad (7)$$

The interpolation parameters α_ℓ and β_ℓ are set to $\ell/(m+\ell)$ for some fixed prior mass m. These estimators have a naturally incremental implementation, either via updating counters or via maintaining sufficient statistics (Duda and Hart, 1973). Furthermore, so long as the meaning of words is fixed over time, then the relative frequencies converge to the probabilities in those formulas. In contrast, network training methods based on total mean-squared error (MSE) involve batch-mode algorithms such as singular value decomposition (SVD) or multipass algorithms such as gradient search (Rumelhart and McClelland, 1988). Observe also that for MSE methods, one must define a smooth distortion measure on the output space, which is not necessary for the information-theoretic network. One can prove however, that for an appropriate error function, the information-theoretic update vector has the same sign components as the gradient of the single-step error (Gorin and Levinson, 1989), i.e., they are moving in the same general direction. One can further prove that the information-theoretic update is guaranteed to decrease that single-step error function (Gorin and Levinson, 1989). A natural next step would be to extend this result to the global error function, but it is not clear how to do so.

Small sample artifacts are an ubiquitous issue in statistical language models. Zipf's law is a well-known empirical observation which tells us that, in general, there will be many low-frequency events and only a few high frequency events (Pierce, 1961) (Zipf, 1949). There are methods which attempt to ameliorate this problem, such as Good–Turing estimators (Good, 1953), used by Rose to estimate mutual information in his topic spotting experiment (Rose *et al.*, 1991).

Due to issues of small sample statistics and context dependency, we are led to investigate *focused learning*, where one would like to adjust the learning rate for a word based on its context. We illustrate this issue with an example from the Department Store system. Consider an input sentence *I want to buy an etagere*, where the word *etagere* is encountered for the first time. Given that the appropriate action is to connect the caller to the furniture department, one should greatly strengthen the association between *etagere* and that action. In contrast, consider a second input sentence *I want to buy a mauve sweater*, where the word *mauve* is encountered for the first time and the appropriate action is to connect the caller to the clothing department. In this example, however, one should learn only a mild association between *mauve* and that call-action. To summarize, depending on context, one would like to accelerate the learning rate for some words, decrease it for others.

We consider how to quantify and exploit this intuition, following (Tishby and Gorin, 1994). Observe that *etagere* is the only possible explanation for the interpretation of the first sentence, while *mauve* is not necessary to correctly interpret the second sentence. That is, the error in understanding is large in one case, small in the other. Thus one would like to modulate the learning rate based on the error, a well-understood principle in MSE optimization algorithms. Tishby observes that there are both algebraic and statistical structures on the network's parameters (Tishby and Gorin, 1994). The algebraic properties can be addressed via MSE methods and the statistical properties via relative frequencies. He combines these, proposing an algebraic method for estimating statistical associations, often obtaining good estimates for words which occur only once.

Farrell investigates a gradient solution to the algebraic formulation, achieving similar performance to the information-theoretic network on the Department Store task (Farrell *et al.*, 1993). That work addressed only the algebraic formulation, not considering the statistical structure. Geutner *et al.* (Geutner *et al.*, 1993) describe a hybrid approach leading to improved results, using mutual information as an initial estimator followed by MSE optimization. We conjecture that this result can be explained via the dual structure on the parameter space. We close this discussion by observing that, while very promising, this line of thought on exploiting algebraic structure to improve statistical associations remains an open research issue both theoretically and empirically. In Sec. VII we present an alternate approach to focused learning, reporting on preliminary experiments in exploiting estimated syntactic state to adjust learning rate.

It can be shown that when words or phrases have uniformly weak associations, then the estimates of their connection weights have increased variance, thereby injecting additional noise into the understanding process. This leads us to consider clipping to zero the weights of those words with weak associations. In (Gorin *et al.*, 1991), this was implemented by clipping the estimates of $P(c|v)$ to $P(c)$ if they were sufficiently close. We have recently introduced an improved method to address this problem, clipping the connections of low-salience words to zero. This *salience thresholding* both reduces the effective vocabulary and increases the understanding rate, as discussed later in Sec. V. Such subvocabulary selection is of great relevance to designing and evaluating the speech recognition front-end of our systems.

Summary. In this section we have described the information-theoretic connectionist network[1] which is the basic building block of our language acquisition systems. Several observations are in order. First, in all of our experiments, the vocabulary and parameter space grow over time as new words are encountered. Second, the networks are embedded in a dialog control system, adapting their parameters based on reinforcement feedback from the environment. Third, as described in Secs. IV and VII, this basic network is embedded in larger structured networks to enable language acquisition for more complex devices.

II. DIALOG CONTROL

We govern the behavior of a device based on feedback as to the appropriateness of its actions. Such reinforcement feedback causes an immediate modification of the device's behavior (*control*) and then a modification of the device's future behavior (*learning*).

In our communication paradigm, input is provided by a person, whose goal is to induce the machine to perform some action. The interaction between human and machine is called *dialog*, which serves the important role of resolving ambiguities and misunderstandings. This interaction between the machine and its environment is implemented as a feedback control system, as was illustrated in Fig. 1. In this section, we describe the basic dialog control mechanism used in our systems.

The initial input to the system is a natural language request for the machine to perform some action. Based on the machine response, the user responds in turn with a mixture of error feedback plus possibly clarifying information. Examples of such dialogs will be provided for each of our experimental systems in subsequent sections. In this section, we describe the formulas underlying the most basic system, then comment upon its properties and extensions.

Let s_ℓ denote the ℓth user input, $\mathbf{a}(s_\ell)$ the activation vector produced by the network [cf. formula (4)], and $\mathbf{e}(s_\ell)$ (for $\ell \geqslant 2$) the error component of the message. Let c_{k_ℓ} denote the machine response after the ℓth input message. Define a total activation array at each stage of the dialog via

$$\mathbf{A}_1=\mathbf{a}(s_1) \tag{8}$$

$$\mathbf{A}_\ell=(1-\alpha_\ell)\mathbf{A}_{\ell-1}+\alpha_\ell\mathbf{a}(s_\ell)+\mathbf{e}(s_\ell). \tag{9}$$

In the simplest case (Gorin *et al.*, 1991), we set the components of $\mathbf{e}(s_\ell)$ to zero, except for $e_{c_{k_{\ell-1}}}(s_\ell)$ which is set to $-\infty$. The most basic implementation (Gorin *et al.*, 1991) set $\alpha_\ell=1/\ell$, so that $\mathbf{A}_\ell$ involves an average of the terms $\mathbf{a}(s_t)$, $1\leqslant t\leqslant\ell$. The components of $\mathbf{A}_\ell$ are denoted $A_{\ell k}$, and the machine's response after the ℓth input involves the action c_{k_ℓ} given by

$$c_{k_\ell}=\arg\max_k A_{\ell k}. \tag{10}$$

Extensions. This basic algorithm has undergone several extensions. In our earliest experiments (Gorin *et al.*, 1991), reinforcement feedback was provided only by the user. Miller showed how to internally generate reinforcement feedback using confidence models (Miller and Gorin, 1993a,c). For systems with multidimensional action spaces, Miller also described how to focus the reinforcement feedback on one or more of the semantic primitive actions (Miller and Gorin, 1993c). In systems with multisensory input, Sankar showed how to combine reinforcement from the user and environment (Sankar and Gorin, 1993). For long dialogs where a sequence of actions is desired, Sankar presented an algorithm which segments long dialogs into regions of stationary intent (Sankar and Gorin, 1993). These issues will be illustrated in the examples of Sec. III, and the interested reader is referred to Miller and Gorin (1993a,c) and Sankar and Gorin (1993) for further details.

Convergence. In evaluating a language acquisition system, the primary measure of performance is how often a device responds appropriately to an input stimulus. We expect, however, our systems to make occasional errors, especially when encountering unfamiliar input, leading to an emphasis on error detection, recovery and learning. This perspective led us to further evaluate how many interactions are required for the user and machine to negotiate an appropriate machine action, which we call *dialog length.* In Gorin *et al.* (1991), a simple dialog convergence model was presented. It predicts that the probability of a dialog converging within L interactions approaches unity exponentially fast with L. Empirical evidence supporting that convergence model was reported for the Almanac (Miller and Gorin, 1993c) and Blocks World (Sankar and Gorin, 1993) systems.

III. APPLICATION TO AUTOMATED CALL ROUTING

In this section, we describe experiments which involve the single-layer information-theoretic network. In particular, consider the task of Automated Call Routing in a telecommunications network. The goal is that when a person desires some service, he would dial a single universal number, which prompts him with *Hello, how may I help you?* He responds with unconstrained fluent speech, upon which basis the call would be routed to an appropriate destination. This scenario can be contrasted with current methods of providing separate telephone numbers for each service or of requiring people to navigate a menu-driven system. In this new scenario, a call would instead be switched on the *basis of its content.*

The Department Store. Our first experiment in this direction involved a Department Store scenario (Gorin *et al.*, 1991). There were three departments: Furniture, Clothing, and Hardware, plus a fourth call-action where the device "gives up" and connects the call to an operator. Perhaps the best method of exposition is to examine several illustrative human/machine dialogs.

Session One
Hello, this is the J.C. Bell Department Store.
How may I help you?
I need some paint for my redwood table.
I will connect you to the hardware department?
Okay.
Please hold on, I am connecting you to the hardware department.

In Session One, the device correctly understood the input, having over time learned to balance the associations of *table* and *paint*. The second input, *Okay*, is a positive reinforcement of the proposed machine action.

Session Two
Hello, how may I help you?
I'd like to buy an etagere.
Maybe should I connect you to the clothing department?
No, it's a kind of furniture.
I will connect you to furniture department?
Okay, that's better.
Please hold on, I am connecting you to the furniture department.

In Session Two, the device misunderstood the initial request because it has not previously encountered the word *etagere*. The second input involves the word *no*, as negative reinforcement, plus a clarifying message.[3] The dialog then converged, leading to the appropriate call-action and network adaptation. The new word is added to the input layer and all connection weights are updated, as was described in Sec. I.

There is a rudimentary confidence model in this early system, reflected in the machine's use of the word *maybe*. Observe that explicit positive reinforcement is necessary before the call was actually routed. Improvements to this basic dialog controller were introduced in subsequent systems, as discussed later in this section. A subsequent dialog in Session Three again involves *etagere*, but in a different context, demonstrating the acquisition of the new word and its meaning.

Session Three
Hello, how may I help you?
I'm looking for an etagere.
I will connect you to the furniture department?
Okay.

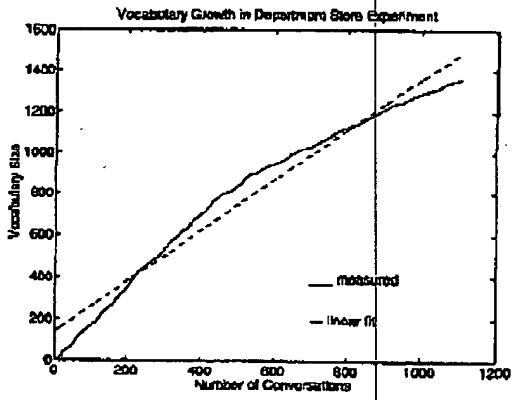

FIG. 3. Vocabulary growth in the department store call router.

Please hold on. I am connecting you to the furniture department.

An experimental evaluation of the Department Store system with keyboard input was reported in Gorin *et al.* (1991) and with spoken input in Gorin *et al.* (1994a). We postpone discussion of the speech system to Sec. VI, focusing here on the system with keyboard input. Initially, the system knows *nothing* about the vocabulary for its task. It was provided with the concepts of word, phrase, and sentence, but *no* instantiations thereof for the task. A word was defined to be any character sequence delimited by blanks or punctuations, a phrase to be any adjacent pair of words. The machine was initially provided with the words *no* and *okay* plus their associations to negative and positive reinforcement, respectively. An experiment was conducted in which 12 users (colleagues at Murray Hill) interacted with the system over a two month period in a total of 1105 dialogs. The vocabulary growth due to the first input sentence of each dialog is shown in Fig. 3.

It is illustrative to examine the *association vector* for several words, as shown in Table I. To each word, there is a three-component vector comprising the word's mutual information with the call-actions: routing to the *F*urniture, Clothing, or *H*ardware Departments. Those words are selected for illustration, rather than in any particular order.

As expected, the word *sweater* has strong positive associations to clothing, else negative. Similarly, the word *glue* has strong positive associations to hardware. The word *need* illustrates an interesting usage pattern, where when someone "needs" something, then it is more likely to be hardware.

TABLE I. Network associations in the department store system.

Word	F	C	H
SWEATER	−2.32	+1.34	−2.32
GLUE	−3.17	−3.17	+1.34
NEED	−0.28	−0.50	+0.50
MOTHER	−1.58	+1.19	−1.58

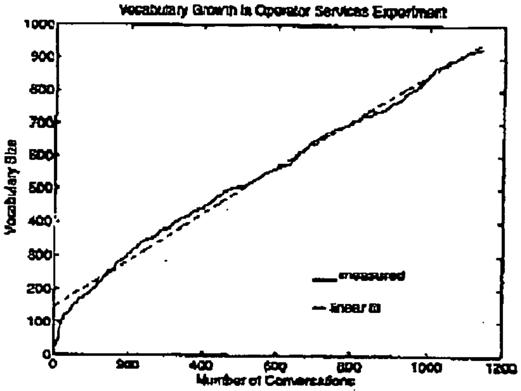

FIG. 4. Vocabulary growth in the operator services call router.

The final example in Table I was brought to our attention during a laboratory demonstration, when a visitor provided the input *I need a birthday present for my mother*, whence the machine confidently offered to connect us to the clothing department. In response to the accusation of constructing a politically incorrect device, we could only respond that this was not preprogrammed: the machine is just a product of its environment.

Operator Services. Based on the above experiment, we became interested in how such methods would apply to real-world data. To this end, a small speech database was collected of actual customer/operator transactions in the AT&T network. The customer input was orthographically transcribed and labeled with one of 21 call-actions. The vocabulary growth over the course of 1140 input sentences is shown in Fig. 4, growing to 932 words. Some of these words are significant for the task, others not. A discussion of this issue is postponed to Sec. V, where we quantify the notion of *salience* and include tables of the most salient words for this and other tasks.

To each word, there corresponds a 21-component association vector. One interesting example is the word *HOME*, which is strongly associated with the call-action of third-number billing, typically embedded in an input such as *I want to charge this to my home phone please*. Another example is the word *CHARGE*, which is positively associated to the two call-actions of third-number and credit-card billing.

Sankar reported on preliminary experiments in mapping these transcriptions to call-actions (Sankar *et al.*, 1993). An on-line conversational-mode system has been constructed for this task, initially with keyboard input (Miller and Gorin, 1993a), then with spoken input (Gorin *et al.*, 1994a) using the methods of Rose (1993). The dialog controller is improved over that of the Department Store system, exploiting a confidence model as illustrated below. In Session One, the machine encounters an ambiguous input, which is resolved via reinforcement feedback plus clarifying input. In Session

Two, the machine has high confidence in its understanding of the input, proceeding directly to the call-action without waiting for explicit external confirmation.

Session One
Hello, how may I help you?
I want to charge this call.
You want to charge this call to a third-number?
No, to put it on my credit card.
You want to charge this call to your credit card.
Please enter your card number:

Session Two
Hello, how may I help you?
I need an emergency operator right now.
Please hold on while I connect you to an operator.

We conclude this section by commenting on other tasks involving sorting natural language into classes or bins. There have been several recent experiments involving topic identification from speech (McDonough *et al.*, 1994) (Peskin, 1993) (Rohlicek, 1992) (Rose *et al.*, 1991). The task of sorting and routing text data has been addressed by many researchers (Belkin and Croft, 1992; Geutner *et al.*, 1993). These tasks differ from the Call Routing application in several dimensions. First, in Call Routing, the input is provided by a cooperative user who desires to be understood and who is cognizant of the general range of call-actions. Second, the input is typically one sentence, rather than whole paragraphs or conversations. Third, there is the opportunity for dialog with the user, serving to resolve ambiguities and misunderstandings.

IV. STRUCTURED NETWORKS

As a device and its environment become more complex, so does the mapping from input stimuli to action. In order to provide rapid learning and generalization in a language acquisition device, we have proposed utilizing structured networks which reflect the device's input/output periphery and environment. The network structure provides functional constraints on the mapping from stimuli to action, thereby greatly accelerating the learning process. We have developed several methods for constructing large structured networks from component subnetworks, experimentally evaluating these ideas in the Almanac and Blocks World systems in this section, and in the Airline Information system discussed in Sec. VII.

A. Product networks and the Almanac system

In the Call Routing experiments of the previous section, the set of machine actions comprises a simple list. Miller investigated the situation where the set of actions comprises an n-parameter family of subroutines (Miller and Gorin, 1993c). The individual selection of parameter values are denoted *semantic primitive actions* for the device. If the n parameters are common to all the subroutines, then the action space is isomorphic to the Cartesian product of the semantic primitive actions. Miller then proposes the construction of a product network, where individual networks are assigned to the selection of each semantic primitive value.

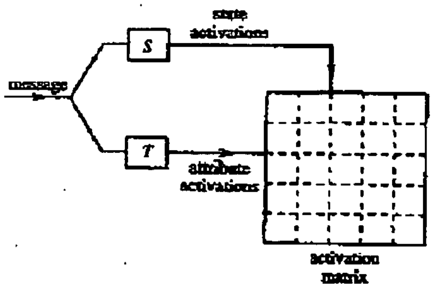

FIG. 5. Product network in the Almanac data retrieval system.

A two-dimensional product network was evaluated in an Almanac data retrieval task. In particular, the device knows 20 facts concerning each of the 50 States (in the U.S.A.), and is capable of retrieving any of these 1000 facts. For example, given the input *What is the capital of New Jersey?*, then an appropriate response would be *The capital of New Jersey is Trenton*. As illustrated in Fig. 5, two information-theoretic networks are combined in a product network for this task. The input is applied to both networks independently, then the resultant activation arrays combined in an outer sum as follows. Denote by $\{a_1(n)\}$ $1 \leqslant n \leqslant 20$ the activation array produced by the attribute-selection subnetwork, by $\{a_2(m)\}$ $1 \leqslant m \leqslant 50$ the array produced by the state-selection subnetwork. Let $\oplus$ denote the outer sum of two vectors, producing a matrix whose entries are pair-wise sums of the vector components. In particular, we compute the 20 by 50 activation matrix via

$$\langle a_{nm}\rangle = a_1 \oplus a_2 = \langle a_1(n) + a_2(m)\rangle. \qquad (11)$$

The action is then performed corresponding to the ordered pair with maximum activation. We remark that for processing the first utterance in a dialog, it would suffice to select the maxima separately for each dimension. As detailed in (Miller and Gorin, 1993c), however, the matrix representation becomes useful in later stages of the dialog, where one must interpret negative reinforcement and estimate confidence models.

We describe this system via several illustrative dialogs, as follows.

Session One
HELLO. I KNOW FACTS ABOUT THE FIFTY STATES. WHAT DO YOU WANT TO KNOW?
How big is Iowa?
WHAT DO YOU WANT TO KNOW ABOUT IOWA?
How much land is there?
IOWA'S TOTAL LAND AREA IN SQUARE MILES IS 56,275.

Session Two
WHAT ELSE DO YOU WANT TO KNOW?
Do you know how big New York is?
DO YOU WANT TO KNOW THE TOTAL LAND AREA OF NEW YORK?

Yes.

NEW YORK'S TOTAL LAND AREA IN SQUARE MILES IS 49,108.

In Session One, the machine was unfamiliar with the phrase *How big is...* Negative reinforcement was generated internally via a confidence model, leading to the machine responding via a partial prompt for the attribute. A subsequent dialog was shown in Session Two, demonstrating the acquisition of those words and their meaning for this task. Observe that the product network has enabled generalization of that meaning to a new state (New York), although no explicit supervision was provided as to which words are associated to which semantic primitive action. We now illustrate how dialog control can resolve an ambiguous input.

Session Three

WHAT ELSE DO YOU WANT TO KNOW?
What's the main city in California?
DO YOU WANT TO KNOW THE LARGEST CITY OF CALIFORNIA?
No, its capital.
DO YOU WANT TO KNOW THE CAPITAL OF CALIFORNIA?
Yes.
SACRAMENTO IS THE CAPITAL OF CALIFORNIA.

Session Four

WHAT ELSE DO YOU WANT TO KNOW?
What's the main city in Texas?
DO YOU WANT TO KNOW THE CAPITAL OF TEXAS?
Yes.
AUSTIN IS THE CAPITAL OF TEXAS.

In Session Three, the machine misunderstood the intent of the ambiguous phrase *What's the main city* ... . In the second input, the user provided negative reinforcement (*no*), plus clarifying information. Observe that the device decides for itself whether the user is rejecting its understanding of the state or attribute value. Session Four demonstrates the acquisition of meaning for that phrase and its generalization to a different state.

An experimental evaluation was conducted of the Almanac system with keyboard input, involving 13 users over a two week period in a total of 1018 dialogs (Miller and Gorin, 1993a). Figure 6 shows the vocabulary growth due to the initial input sentences over the course of these dialogs.

B. Sensory primitive subnetworks and the Blocks World

In many situations of interest, the appropriate machine response depends not only on the spoken input but upon the state of the environment. We are thus motivated to investigate devices with multisensory input, which learn the mapping from spoken input *plus* the state of the world to an appropriate machine action.

Consider a robotic device commanded to *Pick up the blue cube*. The appropriate machine action then depends on where the blue cube is actually located, which must somehow be sensed. For a second example, consider the Automated Call Routing task described in Sec. I. For an input such as *Help! This is an emergency!*, the appropriate call-destination depends on the physical location of the telephone from which the call was initiated. A third example, from a teleconferencing control task such as Humanet (Flanagan *et al.*, 1991), is the spoken command *Lights, please*, to which the appropriate action depends on whether the lights are currently on or off.

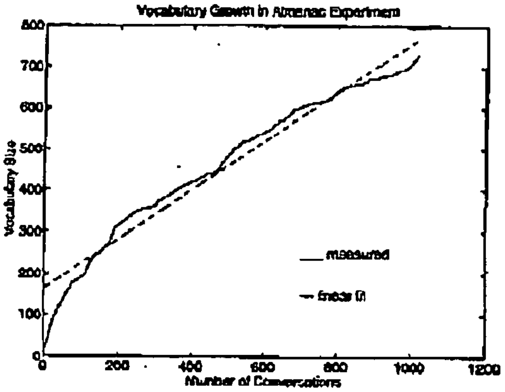

FIG. 6. Vocabulary growth in the Almanac system.

Sankar investigated adaptive language acquisition in a multisensory *Blocks World* (Sankar and Gorin, 1993). The machine was presented with a simulated visual scene, containing several objects of different colors and shapes. In response to input such as *Where is the red square?*, the device demonstrates its understanding via focusing its "eyeball" on the appropriate object. The action space of this device is parameterized by a two-dimensional continuum of eyeball coordinates, specifying the device's *visual focus of attention.*

The device was provided with several innate characteristics, implemented via a time-varying potential function. First, it can sense the color and shape of the objects in its visual scene and it is attracted to bright or moving objects. Second, after focusing on some object, it becomes bored and its attraction to that object diminishes over time. Third, it also constructs associations between linguistic and visual events that co-occur temporally, and is then attracted to objects which are strongly associated with its linguistic input.

Sankar proposed the construction of *sensory primitive subnetworks* which learn associations between the linguistic and visual sensory inputs, implementing the third characteristic above. The output of these subnetworks are then combined via a product network, yielding a time varying potential function over the visual scene. The eyeball motion is then governed by directing it towards the minimum of that potential function. Figure 7(a)–(d) illustrate a sequence of interactions with the system, where it acquires the meaning of the word *circle*. This conversational mode system, with keyboard input, was evaluated via interaction with eleven

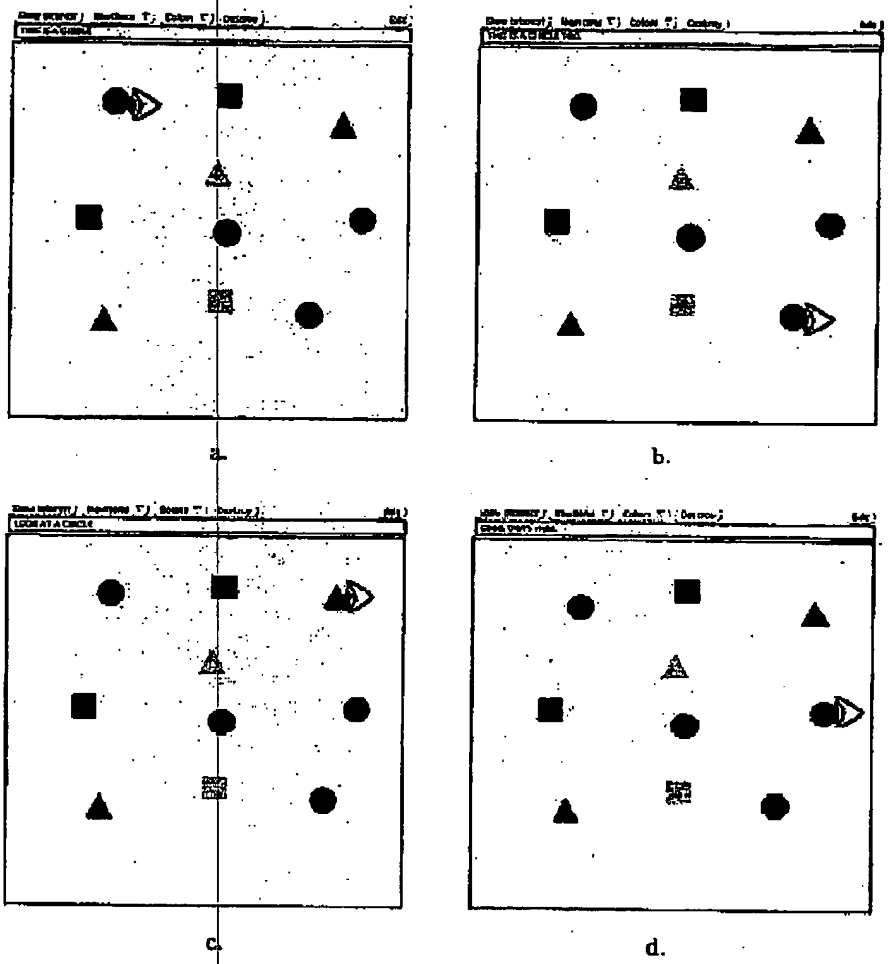

FIG. 7. Language acquisition in a Blocks World.

users in over 1000 unconstrained natural language dialogs, acquiring a vocabulary of 131 words.

Based on the principle that language acquisition for complex devices should proceed in developmental stages, we observe that visual focus of attention is a prerequisite step to devices which actually manipulate the objects in their field of view. Henis reports on an extension of Sankar's experiment, connecting it to a robotic simulator and demonstrating language acquisition for several manipulatory actions with six sensory input channels (Henis *et al.*, 1994).

In this section we have discussed methods for building structured networks from component subnetworks, which have been experimentally evaluated in an Almanac and a Blocks World System. We now investigate how to exploit such a network of associations to quantify the meaning and information content of language.

V. MEANING, SALIENCE, AND INFORMATION

We have been investigating devices which learn to understand and act upon spoken input. The ultimate goal of these speech understanding systems is to extract meaning from the speech signal. A crucial issue in engineering such devices is to quantify the information content of spoken natural language, then to measure a machine's success in extracting that information.

While information theory is a well-developed discipline, there is the standard caveat that its notion of information is quite different from a layman's. In contrast, for systems which understand spoken language, the *semantic* aspects of communication are *primary*. In this section, we propose principles and mechanisms for quantifying the semantic attributes of language. In this and later sections, we discuss the implications of these methods for vocabulary selection in wordspotting, acquisition and adaptation of new spoken words, grammatical inference and robust parsing.

A. Semantic/sensory associations

For people, an input stimulus evokes memories of associated perceptions and activities. We are thus motivated to define the *meaning* of a word, for a particular device, to be its *network associations* to the device's input/output *periphery*.[4] Such a definition grounds meaning in a device's interaction with its world, being dependent on its input/output periphery, environment and experiences. We now describe illustrative examples of this connectionist representation of meaning in several experimental systems.

In the simplest case of a single-layer network, each word-node is connected to each output node, so that its network associations comprise an n-dimensional vector (where n is the number of output actions for the device). For the Department Store Call Router described in Sec. III, there is a three-component vector for each word, as was illustrated in Table I. For the Operator-Services Call Router, also introduced in Sec. III, the network associations comprise a 21-component vector. An illustrative example from that task is the association vector for the word *charge*, which is most strongly associated with *third-number billing*, has mild associations to *credit-card billing*, and is inhibitory of the other call-actions.

Miller's Almanac data retrieval system is based on a product network architecture, as was described in Sec. IV. In this case, the associations between words and actions are *indirect*, being factored through the semantic primitives of attribute and state-selection. For a product network, the word/action network associations are fully determined by the associations between words and primitives. An illustrative example from that system is the word *Colorado*. Its associations within the state-selection subnetwork are quite predictable, being strongly associated to one state and highly inhibitory of the others. While one might expect *Colorado* to be null for the attribute-selection subnet, it turns out to be moderately associated to requests for highest mountains (not surprising, in retrospect). For Oz fans, it is interesting to observe the strong association between the word *Dorothy* and queries about Kansas. We recall that these associations were *acquired* by the device during its interactions with many users over a period of time.

In Sankar's Blocks World experiment, there are network associations between a word and the visual input periphery, factored through the color and shape sensory primitives. In Henis's extension of that Blocks World, the associations involve six visual features and three machine actions. One could argue that as a device's input/output periphery becomes more anthropomorphic, so will this representation of meaning. This will remain a conjecture, however, until we can construct sufficiently complex devices to test the hypothesis.

B. Salience

Given the representation of meaning via network association vectors, then a *semantic distortion* measure can be introduced between such vectors. A *null word* for a device is one whose network associations are all zero. The *salience* of a word for that device can then be defined via its distortion from the null word, thus providing a "norm" on the network associations vectors.

We now focus our attention on a single-layer information-theoretic network and quantify these intuitions. In this case, the network associations of a word comprise a vector of mutual informations between that word and the various machine actions. There are many possible distance measures that one could explore, but it is advantageous to exploit the information-theoretic nature of the vectors. Given a word, v, denote its network association vector as $\langle I(v,c_k)\rangle$ where $\langle\cdots\rangle$ denotes a vector whose kth component is $I(v,c_k)$. We define a semantic distortion measure $d_m(v_1,v_2)$ between two words v_1 and v_2 via first computing the difference between these vectors, then converting that difference into a scalar via projection onto some vector $\bar{u}$. Denote $\bar{w}_1=\langle I(v_1,c_k)\rangle$ and $\bar{w}_2=\langle I(v_2,c_k)\rangle$, then define

$$d_m(v_1,v_2)=(\bar{w}_1-\bar{w}_2)\cdot\bar{u}. \tag{12}$$

In particular, if we select $\bar{u}$ to be the vector of *a posteriori* probabilities $\langle P(c_k|v_1)\rangle$, then

$$d_m(v_1,v_2)=\langle I(v_1,c_k)-I(v_2,c_k)\rangle\cdot\langle P(c_k|v_1)\rangle. \tag{13}$$

In addition to its geometric interpretation as the scalar projection of a vector-difference, this distortion also has an information-theoretic interpretation. It can be easily seen that formula 13 is equivalent to the Kullback–Leibler distance (a.k.a. relative entropy) between the *a posteriori* distributions $\langle P(c_k|v_1)\rangle$ and $\langle P(c_k|v_2)\rangle$ (Cover and Thomas, 1991). That is, the semantic distortion between the two words is equivalent to the distance between the distributions that they induce on the network's periphery.[5]

Recall that a null word, v_{null}, is one whose association vector is all zeros. The *salience* of a word for a given device is defined as its semantic distortion from v_{null},

$$\text{sal}(v)=d_m(v,v_{\text{null}})=\sum_{k=1}^{K} P(c_k|v)I(v,c_k). \tag{14}$$

It was shown by Blachman that this is the unique nonnegative measure of how much information a value of one random variable provides about a second one (Blachman, 1968). That is, denoting the random variable of machine actions by C, then sal(v) is a measure of how much information the word v provides about C. Thus, salience provides an information-theoretic measure of how meaningful a word is for a particular device. After describing a few examples of this measure, we will empirically distinguish it from Shannon's measure of the information content of a word.

Illustrative examples of the most salient words for the Operator Services task are given in Table II. The primary call-action is the one with maximum association to that word. In those cases where there is a moderately strong secondary association, then it is also included. The most salient word, *HOME*, almost always occurs in a single call-action, for example in phrases such as *bill this to my home phone*.

TABLE II. Salient words in the operator services task.

Word	Salience	Primary call-action	Secondary call-action
HOME	2.57	third-number billing	
CARD	2.27	calling-card billing	
CREDIT	2.19	calling-card billing	wrong-number credit
CHARGE	2.01	third-number billing	calling-card billing
BILL	1.70	third-number billing	calling-card billing
AT&T	1.64	switching carriers	calling-card billing
CODE	1.36	area code request	third-number billing
PHONE	1.30	third-number billing	calling assistance
CALLING	1.29	calling-card billing	person-to-person billing
MY	1.11	third-number billing	calling-card billing
COLLECT	1.06	collect call	
FROM	0.69	collect call	

Sankar showed that one can select a subvocabulary for this Call Routing task via salience-thresholding, i.e., clipping the connection weights of low-salience words to zero (Sankar *et al.*, 1993). In that experiment, it was shown that this weight-clipping actually improved the understanding rate from transcriptions. We conjecture that this result is due to the increased variance of the weight estimates for low-salience words. More recently, we have reported on analogous results when applying the network to the output of a continuous speech recognizer (Gorin *et al.*, 1994b). One might also consider exploiting such methods for reducing the vocabulary of a speech recognizer in order to produce a wordspotting system. Related issues have been addressed in (McDonough and Gish, 1994) (Peskin *et al.*, 1993) but remain subjects for future research.

The previous example was for a device with a single-layer network. For devices with more complex network architectures and peripheries, such as the Almanac and Blocks World, we can separately measure the salience of words for each of their semantic and sensory primitive subnetworks. Table III contains some salient words for the attribute-selection subnetwork of the Almanac system. Table IV shows some of the most salient words for the color sensory primitive subnetwork of the Blocks World system, where the multilingual content is striking.

TABLE III. Salient words in the almanac data retrieval task.

Word	Salience
FLOWER	4.39
NICKNAME	4.16
SONG	4.09
DENSITY	4.09
ELEVATION	3.66
CROP	3.65
WHEN	3.60
INDUSTRY	3.59
ENTER	3.56
WHO	3.48
PAY	3.39
BECOME	3.38
TAKE	3.27
MUCH	3.15
BIG	3.13
MURDERS	2.98
LEADING	2.94
LARGEST	2.85

TABLE IV. Most salient words for the color sensory primitive.

Word	Salience
YELLOW	1.96
BLUE	1.93
GREEN	1.81
LIME	1.70
AZUL	1.64
HARA	1.59
RED	1.54
SKY	1.44
MAROON	1.37
ROUGE	1.35
BLOOD	1.35
JAUNE	1.34
GRASSY	1.34
LAL	1.29
PILA	1.28
SOBUJ	1.26
PEELA	1.19
BURNED	1.18
CRIMSEN	0.87

C. Salience versus information

We briefly distinguish and compare the notions of salience and information. The *entropy* of a language measures how much information is produced, on the average, for each symbol in the language (Shannon, 1951). For the sake of exposition, let us consider single words at a time, ignoring issues involving correlations between adjacent words. The information content of a word v is defined as

$$i(v) = -\log_2 p(v), \tag{15}$$

whose average over all words is an approximation of the language entropy. Both entropy and $i(v)$ are measured in *bits per word.*

We first observe that these definitions involve only the language itself. For example, given *War and Peace* in the original Russian, one could compute the information content of individual words and the entropy of the language, without ever understanding a word. In contrast, computing salience involves both the language and its extra-linguistic associations to a device's environment.

It is illustrative to empirically quantify this distinction in the Operator Services task (cf. Section III). First, we compute $i(v)$ for each of the words, using relative frequency estimates of their probabilities. The distribution of the $i(v)$ is shown in Fig. 8, plotted on a log scale. The near-linear form of this distribution is not capricious, and can be related (Gorin *et al.*, 1994a) to Zipf's law (Zipf, 1949). Similarly, we also compute sal(v) for each word, whose distribution is shown in Fig. 9. A scatter plot of salience versus information content for this database is shown in Fig. 10. It is clear that while these measures are quite different, there are relationships between the two that are reflected in the structure of the scatter plot. Characterization and exploitation of this structure is a subject for future research.

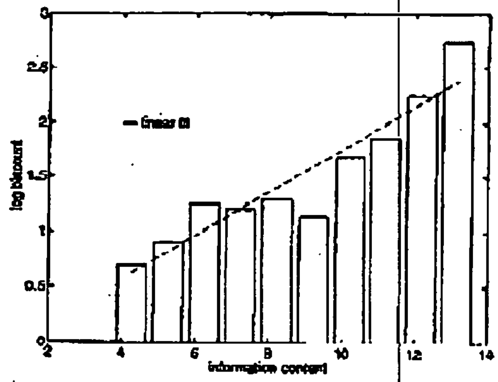

FIG. 8. Information content in the operator services vocabulary.

Summary: In this section, we have explored how to define meaning via semantic/sensory associations, then quantified such intuitions via a salience theory. In the next section, we will comment upon how one might exploit salience in the acquisition and adaptation of spoken words.

VI. EXPERIMENTS IN SPOKEN LANGUAGE ACQUISITION

In traditional speech understanding systems, the symbols and structure in the linguistic hierarchy are defined *a priori*. For humans, however, the symbols of language seem to emerge naturally during the course of a child's interaction with his environment (Kuhl, 1992). This contrast motivates us to investigate how to mimic such behavior in our machines, then how to exploit it to improve the trainability, adaptability, and robustness of our speech understanding systems.

In this section, we report on experiments involving automated acquisition of spoken words, which are the fundamental unit of meaning in language. Albeit rudimentary, the systems are unique in that there is *no* text utilized during training or evaluation, in contrast with all other spoken language systems (Levinson and Shipley, 1980) (Pieraccini, 1992) (Rabiner and Juang, 1993) (Ward, 1991) (Zue, 1992).[6] Our experiments are also unique in that the vocabulary words and their meanings are acquired automatically, in contrast to all other systems where the salient vocabulary is predefined.

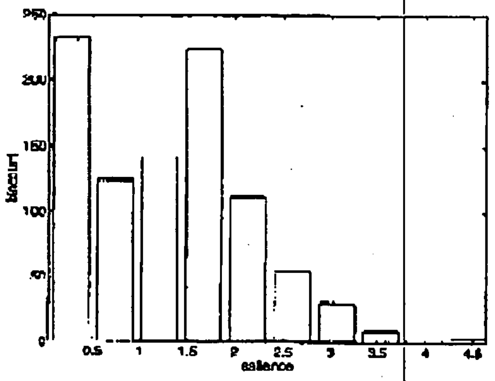

FIG. 9. Salience in the operator services vocabulary.

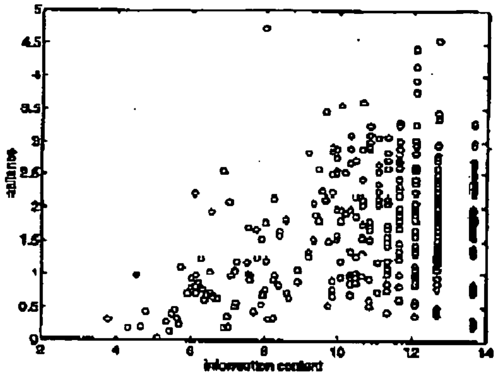

FIG. 10. Salience versus information content in the operator services vocabulary.

The vehicles for these speech experiments are the Department Store and Almanac tasks, which were introduced as keyboard-based systems in Secs. III and IV respectively. In each case, the input was constrained to sequences of isolated spoken words. For each task, we had previously recorded many keyboard dialogs from multiple users for each. The initial input from each dialog was read and recorded by a single speaker in an office environment. These utterances plus their corresponding semantic actions provided the database for these experiments. In the case of the Department Store, the resultant network was embedded into a conversational-mode system with speech input and output, which acquired new words and adapted known ones during the course of performing its task.

A. New word acquisition

The utterances were segmented into individual word tokens via their energy contours (Wilpon *et al.*, 1984). Feature extraction comprised 12 cepstral and 12 delta-cepstral coefficients at 10-ms intervals (Rabiner *et al.*, 1989). As detailed in (Gorin *et al.*, 1994a), there are two stages in the incremental training algorithm for each (utterance,action) pair. First is an adaptive clustering of the new word tokens into existing word-nodes, possibly creating one or more new word-nodes. Word tokens were compared via a Dynamic Time Warping (DTW) measure (Itakura, 1975) with a local Euclidean distance. Second, the information-theoretic connection weights are updated via the methods of Sec. I.

The word-nodes in these experiments are represented via a cluster of spoken word tokens, providing the device with

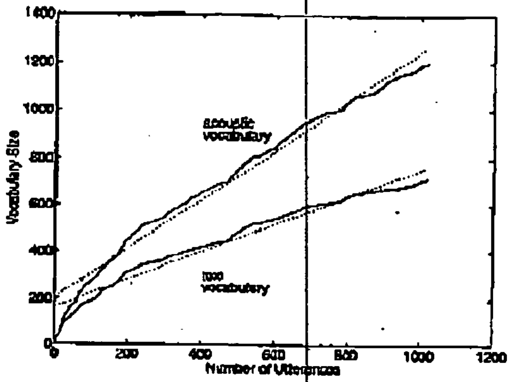

FIG. 11. Acoustic vocabulary growth in the Almanac.

an acquired *acoustic vocabulary*. Since *no* text is provided to the system, it is illuminating to compare this acoustic vocabulary with the corresponding text vocabulary. There are two kinds of "errors" that occur at this level. The first type of error is when word-tokens with different orthography are *merged* into a single acoustic cluster. The second is when word-tokens with the same orthography are *split* into separate acoustic clusters.

Figure 11 illustrates the acoustic vocabulary growth for the Almanac experiment (Miller and Gorin, 1993b), compared with the corresponding text-vocabulary growth. It is clear from this plot that the split phenomenon dominates, at least for this particular clustering algorithm and parameter settings. The same situation is observed in the spoken Department Store experiment (Gorin *et al.*, 1994a).

We quantify the split phenomenon by computing the number of acoustic clusters per text-word, which is histogrammed in Fig. 12 for the Almanac task. Observe that most words are in only one or two clusters. Most of these splits are observed for words such as *a, the, in, what,* and *state* (Miller and Gorin, 1993b). One can also analyze the merge-phenomenon, yielding a similar distribution with an average of 1.15 text-words per acoustic cluster. Table V provides an analysis of these merged words for the Almanac system. Observe that more than half are due to misspellings and abbreviations in the text—i.e., errors in name only, not in fact. The first row in Table V is partially an artifact of "read speech," where text such as *NV* and *Nevada* were pronounced identically. Table VI provides, for the Department Store task, a listing of various acoustic clusters which contain more than one text-word. Observe that some of these merges are semantically significant for the task (e.g., *bread/red*) while others are not (e.g., *nail/nails*).

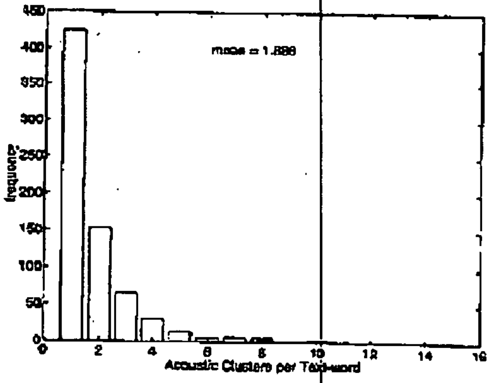

FIG. 12. Distribution of split words in the Almanac task.

TABLE V. Analysis of merged words in the Almanac acoustic vocabulary.

Classification	Examples	Percent
Misspellings and abbreviations	peole, people NV, Nevada	56%
Monosyllabic	in, is, it, tis date, state for, or, poor great, rate	23%
Plurals and suffixes	state, states name, named	13%
Homophones	main, Maine road, Rhode	4%
Other	be, being, the	4%

B. Joint acoustic/semantic distortions

These split/merge phenomena are quite sensitive to threshold-parameter values in the adaptive clustering algo-

TABLE VI. Some examples of merged words in the department store system.

at that	bed head	bit get	blue glue	bread red	chain jean
does doesn	find fine	free three	lists that's	hello pillow	hose those
i'm um	know no	love of	many mini	nail nails	new on
piece these	saw so	the time	wants wines	watt what	wear where
wall will that	wood would	at hat	beds it's its	floor for four	hair pair tear
lick light like	made make need	stain stained stand	air pair there they're	a ahh of the	buy buying by dye hi
do to too two	gas guess kiss quess	it kid kit yet	sale sell so still	an and been end in	high i i'd my

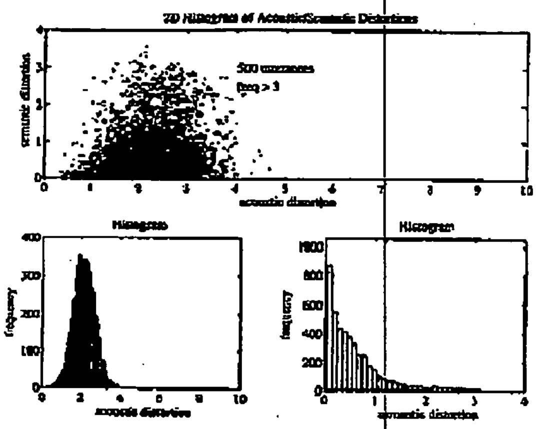

FIG. 13. Joint acoustic/semantic distribution in department store vocabulary.

rithm, as discussed in (Gorin *et al.*, 1994a). Presumably, with much improved feature analysis and distortion measures, the acoustic and text lexicons should become nearly identical (modulo homonyms and homophones). For humans, however, we observe that our perceptual metric is a partially acquired trait. That is, there are sound distinctions which are perceived (or not) depending on where an individual was raised (Kuhl, 1992).

This motivates us to investigate how to exploit nonacoustic information to improve the acquisition and adaptation of spoken words. Although a topic for future research, we make some preliminary observations in this direction. In particular, one could investigate clustering methods which acquire spoken words in the composite acoustic/semantic space. For each pair of text-vocabulary words in the Department Store task, we compute their *acoustic distortion* via a minimum nearest-neighbor distance between the spectral templates in their respective clusters. Then, for each such word-pair, we also compute a symmetrized semantic distortion using the methods of Sec. V,

$$\bar{d}_m(v_1,v_2)=\tfrac{1}{2}[d_m(v_1,v_2)+d_m(v_2,v_1)]. \qquad (16)$$

A scatter plot of these two distortions is shown in Fig. 13, along with projections onto the individual axes. The acoustic distortion distribution is essentially gaussian, as discussed further in Gorin *et al.* (1994). The semantic distortion distribution is exponential, and is nearly linear if examined on a log scale.

In the experiments of Gorin *et al.* (1994a) and Miller and Gorin (1993b), a purely acoustic criterion was used for adaptation and acquisition of spoken words, corresponding to a vertical decision boundary in the two-dimensional histogram of Fig. 13. How might one exploit semantic-level supervision via this joint distribution to improve the acquisition process? We conjecture that a tilted decision boundary might separate semantically distinct words which are acoustically similar (cf. Table VI). While an intriguing possibility, this remains a subject for future research.

VII. GRAMMATICAL INFERENCE

In the previous section, we described some intuitions and experimental results involving automated acquisition of spoken words. The next level up in the linguistic hierarchy (Levinson, 1985) is grammar, traditionally viewed as comprising the rules which constrain how words are put together into sentences (Winograd, 1983). Grammatical structure also modulates the semantic associations of a word, adjusting its meaning depending on where it appears in a sentence. For example, within the Department Store Call Router, the network associations of the word *chair* should be quite different in *I want to buy a chair* versus in the sentence *I need some glue to fix my chair*. In the experiments described thus far, the network associations of a word or phrase have been context-independent. Hence, a natural next step is to extend these methods to encompass context-*dependent* associations of a word when it appears in a particular part-of-speech. An intimately related question is how such parts-of-speech might be automatically acquired. In this section, we recount some intuitions and preliminary experiments in these directions.

There has been much debate over the years over how much of linguistic structure is innate versus acquired (Chomsky, 1965). In particular, grammatical inference from samples of the language has received much attention.[7] However, this is a much more difficult problem than people actually solve, who acquire language not only by listening to it but by using it during the course of interacting with their environment. This contrast motivates us to investigate how to exploit such *extra-linguistic* information in automated language acquisition, governed by the machine's desire to understand and respond appropriately to its input.

We are motivated by human language acquisition, in which the early characterization of a part-of-speech is semantic/sensory, hence grounded in our physical environment.[8] For example, the elementary-school definition of a *noun* is a *"person, place, or thing"* and a *verb* as a word which *connotes action*. One can similarly provide semantic/sensory definitions of adjectives, adverbs, etc. While a subject of debate in linguistics, these definitions underlie our intuition of how to exploit semantic/sensory associations to bootstrap grammatical inference. In particular, we propose to define a part-of-speech, for some device, as a set of words which are strongly associated to some dimension of the device I/O periphery. In formal language theory (Aho and Ullman, 1972), one denotes the vocabulary words as the *terminal* symbols, while parts-of-speech and phrases are described via *nonterminals*. Parts-of-speech are a particular type of nonterminal which describes a word-class, sometimes denoted a *preterminal* (Pereira, 1994).

Our most elementary systems are based on a single-layer network, where the input periphery involves only words and the output periphery is a set of meaningful machine actions. In Sec. V, we quantified the notion of salience, then used it to rank-order the vocabulary for several tasks in Tables II, III, and IV. By selecting a threshold, one can induce a salient preterminal for a task which we denote S. That is, the highly salient words are exemplars of S. One can exploit this preterminal to modulate semantic associations via suppressing the associations of nonsalient words. In particular, denote by $I(v,S;c)$ the context-dependent associations to action c of a word v produced by preterminal S. Denote by $\bar{S}$ the nonsalient preterminal, which is in this case just the complement of S. We can modulate the semantic associations of a word v by defining

$$I(v,S;c)=I(v,c), \tag{17}$$

$$I(v,\bar{S};c)=0, \tag{18}$$

where $I(v,c)$ is the context-independent mutual information described in Sec. III. Sankar showed (Sankar *et al.*, 1993) that this simple modulation leads to much improved understanding rate in the Operator Services task. Furthermore, S can define the subvocabulary for a wordspotting front-end for the task.

In more complex devices, one can induce several preterminals in this manner. For example, in Miller's Almanac data retrieval system, we can induce one preterminal which is salient for state-selection, another which is salient for attribute-selection (cf. Table III). In Sankar's Blocks World system, we can induce salient preterminals for the *color* and *shape* sensory primitives. In Henis' extension of the Blocks World, there are several sensory and semantic primitive layers, leading to a corresponding number of preterminals. Recently, Masukata and Nakagawa (1994) have reported on grammatical inference exploiting speech and visual input in a Blocks World. There are many interesting issues which arise in grammatical inference for such devices. In this paper, however, we report only on some initial experiments for exploiting extra-linguistic associations for grammatical inference in the Call Routing tasks.

A. Trigrams of salient words

Trigrams are an elementary type of grammar, describing the allowable (or probable) adjacent symbol pairs (Jelinek, 1990). There are well-known arguments against the sufficiency of even n-grams on terminal symbols to describe natural language (Chomsky, 1965). Utilizing rules involving trigrams on *nonterminals*, however, is another matter, providing a much more powerful representation (Aho and Ullman, 1972).

We briefly explore this intuition by examining the trigrams of salient words for the Operator Services task (Lee, 1993). Figure 14 shows the left and right context of the word *home*, which is strongly associated with the function of *third-number billing*. The * node indicates a wild-card, i.e. any other word. Observe that *home* is preceded by the word *my* with probability 0.94, and followed by the word *phone* with probability 0.83. This low branching factor is also observed for the other salient words. Such highly constrained local context can be exploited in speech recognition to improve performance of a wordspotting algorithm (Rose, 1993). For example, although it is difficult to spot the monosyllable *home* in fluent speech, it is much easier to spot the phrase "*my home phone*" (Hanek, 1994).

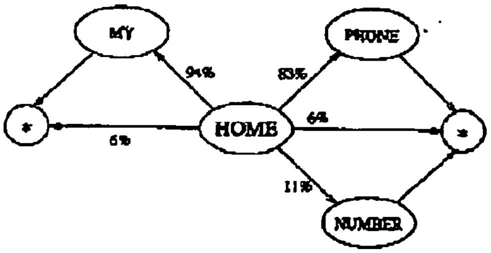

FIG. 14. Local context of the salient word HOME.

Figure 15 shows the context of the salient word *AT&T*, which is slightly more varied. In this example, the context can modulate the meaning of the word. Recall that we have defined the meaning of a word, for some device, to be its network associations to the device periphery. In this particular device, meaning reduces to the vector of associations between the word/phrase and call-actions. For example, when *AT&T* appears in the context " ... *my AT&T card* ... " it should be associated with the call-action of credit-card billing. Alternatively, when it appears in the context " ... *have AT&T long* ... ," it should be associated with accessing AT&T as a long-distance carrier. It remains a subject for future research to experimentally evaluate the utility of these acquired semantic fragments.

B. A finite state grammar with salient states

In natural language, parts-of-speech are characterized both semantically and via within-language usage patterns (Maratsos, 1982). This motivates us to explore evolving our purely semantic characterization of nonterminals in a similar manner. In particular, we report on an early experiment which splits and reestimates salient nonterminals based on within-language patterns.

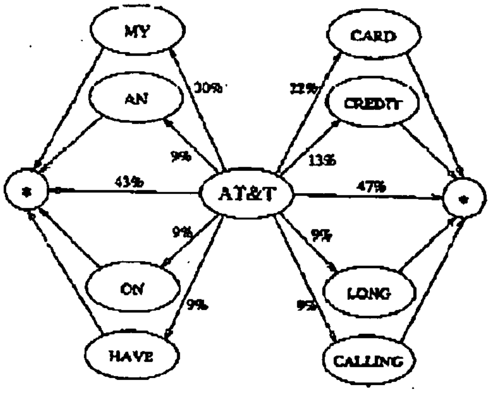

FIG. 15. Local context of AT&T.

TABLE VII. Salient words for the department store system.

Word	Salience
PAINT	1.30
DRESS	1.26
SOCKS	1.21
UNDERWEAR	1.13
SILK	1.13
HAMMER	1.12
HAT	1.10
WRENCHES	1.08
POLISH	1.06
PAIR	1.04
ROOM	1.02
SIZES	1.01
WEAR	0.96
BLACK	0.95
REPAIR	0.81
FIX	0.80
WOODEN	0.73

Table VII shows some of the most salient words for the Department Store task. We split the database into 800 training and 305 test sentences, inducing a preterminal *S* via salience thresholding on the training data. We use Vidal's method of Error Correcting Grammatical Inference (ECGI) (Prieto and Vidal, 1992) (Rulot and Vidal, 1987) to induce a finite state grammar with 18 salient states (Vidal, 1992), a pruned version of which is shown in Fig. 16. Hence, based on within-language context, the single salient state *S* is split and refined into 18 separate states. Five of the states appear in the pruned grammar of Fig. 16, denoted by ⟨*S*⟩. The start-state is indicated by the symbol "~," the stop-state by "@." The graph is directed, with transitions from left to right.

One method of exploiting this grammar would be to modulate network associations depending on whether a word is parsed into a salient state or not, following formulas (17) and (18). To evaluate the feasibility of this idea, a first step is to evaluate whether the grammar does a better job of salience-tagging than merely salience thresholding. We observe that, in the Department Store task, most (64%) of the singleton words are subjectively salient, since participants in the experiment were continually inventing new items to "purchase." To define a salient preterminal, the salience-threshold was thus adjusted to include these words of frequency one. The error rate of that naive salience-tagger on new words is then 36%.

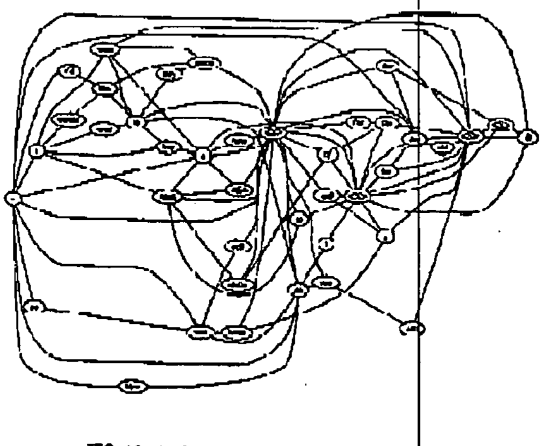

FIG. 16. An induced grammar with salient states.

By way of comparison, the 305 test sentences were analyzed via an error-correcting parser (Prieto and Vidal, 1992) and the complete induced grammar. The words in the test sentences were tagged as salient or not, depending on the state to which they were assigned. In particular, we focus on the dominant salient state, to which 81 new words were assigned. A subjective evaluation showed that only 5 are non-salient, i.e., that the error-rate of the salience-tagger for the dominant state is only 6%, a *sixfold* reduction as compared to the naive salience-tagger described above. While quite preliminary, this experiment indicates the utility of the induced grammar to improve salience-tagging over context-independent salience-thresholding.

This automated salience tagging can also be useful when acquiring new words, for which the estimation of mutual informations is quite noisy (cf. Sec. I). One would like to focus the learning algorithm based on syntactic state. For example, one of the test sentences is *Do you sell* ⟨?⟩ ⟨*S*⟩*flashlights?*, where *flashlights* is a new word (indicated by ⟨?⟩) which is assigned by the parser to a salient state (indicated by ⟨S⟩). Another example is *Do you have* ⟨?⟩ *beanbag* ⟨*S*⟩*chairs?*, where *beanbag* is a new word which is *not* assigned to a salient state. A topic for future research is to provide improved estimators for low frequency words by exploiting such salience-tagging (e.g., accelerating the learning rate for *flashlights* and slowing it down for *beanbag*).

C. An airline information system

Consider a device whose actions involve pairs of places or things. The principle of developmental learning leads us to construct a device which acquires the language for such tasks in stages, first learning the language referring to individual places, only then acquiring the language involving pairs. Gertner constructed a hierarchical network for such tasks, analogous to Fig. 2, except that each node is replaced by a subnetwork corresponding to a linguistic nonterminal (Gertner and Gorin, 1993). The network is hierarchically self-similar, in that each subnetwork is itself a copy of Fig. 2, with nodes corresponding to vocabulary words (terminals) in the language. The subnetworks correspond to nonterminals for place/object names (cf. nouns) and modifier phrases. The construction of associations between phrases and actions is factored through these intermediate nonterminals, providing rapid learning and generalization.

This network was evaluated in an Airline Information Task, a time-honored domain for investigation of natural language understanding systems (Levinson and Shipley, 1980) (Ward, 1991) (Watt, 1968) (Woods, 1968). In particular, the task involved retrieval of a flight table between pairs of cities. We describe the system via several illustrative dialogs as follows (Gertner and Gorin, 1993). The flight tables are not actually retrieved, being illustrated by a sequence of asterisks

(*). Session One illustrates the dialog control system, which resolves an ambiguity in the initial input sentence.

Session One

Machine: HELLO.
I CAN SHOW YOU THE FLIGHT TABLE BETWEEN ANY PAIR OF CITIES.
HOW MAY I HELP YOU?

User: *I'd like to go to Detroit.*

Machine: WOULD YOU LIKE TO SEE THE FLIGHT TABLE FROM BOSTON TO DETROIT?

User: *No, from NYC.*

Machine: WOULD YOU LIKE TO SEE THE FLIGHT TABLE FROM NYC TO DETROIT?

User: *Yes.*

Machine: HERE IS THE LIST OF THE FLIGHTS FROM NYC TO DETROIT.

****

****

Sessions Two and Three illustrate acquisition of modifier phrases that determine which city is the destination and which is the origin of the flight. For the sake of an interesting example, we've chosen the input message to involve the Nepali language, which uses postpositional rather than prepositional modifiers (Yarowsky, 1990). After Session Two, any reasonable network would have memorized the association between that particular sentence and the appropriate action. A product network would even be able to generalize that the phrase *chicago bata* is associated with Chicago being the departure city. One would desire, however, that the system would furthermore learn that these new words, *bata* and *sama*, modify the semantic associations of *any* city name, not just the ones with which they were observed. Session Three illustrates this generalization, demonstrating the efficacy of factoring the word/action associations through the intermediate nonterminals.

Session Two

Machine: ARE YOU INTERESTED IN OTHER FLIGHTS?

User: *Chicago bata Detroit sama.*

Machine: WOULD YOU LIKE TO SEE THE FLIGHT TABLE FROM DETROIT TO CHICAGO?

User: *No, the other way around.*

Machine: I WILL SHOW YOU THE FLIGHT TABLE FROM CHICAGO TO DETROIT.

User: *Yes.*

Machine: HERE IS THE LIST OF THE FLIGHTS FROM CHICAGO TO DETROIT

****

****

Session Three

Machine: ARE YOU INTERESTED IN OTHER FLIGHTS?

User: *Washington sama Boston bata.*

Machine: I WILL SHOW YOU THE FLIGHT TABLE FROM BOSTON TO WASHINGTON.

User: *OK.*

Machine: HERE IS THE LIST OF THE FLIGHTS FROM BOSTON TO WASHINGTON.

****

****

VIII. CONCLUSIONS

We have described the principles and mechanisms underlying our research in automated language acquisition, exploring their utility in several experimental systems involving both text and spoken input. The basic mechanism in these experiments has been an *information-theoretic connectionist* network embedded in a *feedback control system*. This basic network was evaluated in the context of an Automated Call Routing task. For language acquisition in more complex devices, we first introduced the idea of network architectures which reflect the structure of the device's input/output periphery and environment. We then introduced the idea of developmental learning, where a device proceeds from the simple to the complex, from the concrete to the abstract. These ideas were experimentally evaluated in the context of an Almanac Data Retrieval task, a multisensory Blocks World, and an Airline Information task.

We have proposed a *salience* theory which quantifies the information content of spoken language for a particular device, defining meaning in terms of network associations to the device periphery. Preliminary experiments were presented demonstrating the feasibility of exploiting such meaning to improve spoken word acquisition and grammatical inference.

There are several main directions for future research. The first is to demonstrate scalability of these methods to larger tasks such as teleconference facility control, database retrieval and robotic control. The second is to improve performance on "simpler" tasks such as automated call routing and data retrieval, advancing our understanding of how to integrate these methods with wordspotting and large-vocabulary speech recognition. The third direction is to demonstrate how meaning can be exploited to improve the acquisition of robust models of spoken language, in particular at the levels of subword units, words and grammar.

This research forms the basis of a theory of syntax and semantics, where conveying meaning is primary and linguistic structure serves to make such communication robust. Although our experimental devices are thus far rudimentary, we consider them to be the early stages of a long-term investigation into machines which automatically acquire language through interaction with a complex environment.

ACKNOWLEDGMENTS

This tutorial had its genesis in a suggestion by Larry Rabiner, following the special session on Automated Language Acquisition at the 1992 ASA conference in New Orleans. The technical content of this paper is based on a sequence of lectures that I gave at CNET in June 1993, and I am grateful to Christel Sorin for that opportunity to organize this material. I am particularly grateful to Steve Levinson, Louis Auslander, and Juergen Schroeter for encouraging me to go the extra mile and write this paper. I would like to thank several colleagues for their careful review of earlier versions of this manuscript: Hiyan Alshawi, Oded Ghitza, Henry Hanek, Ealan Henis, Fred Jelinek, Steve Levinson, Brian MacWhinney, Laura Miller, Fernando Pereira, Roberto Pieraccini, David Roe, Ananth Sankar, Richard Sproat, Naftali Tishby, and Enrique Vidal. I am also grateful to the anonymous reviewers for their insightful comments. Finally, I thank Martina Sharp for her accuracy, hard work, and patience in creating this manuscript.

[1] The reader is referred to Elman (1991) Fischer (1980), and Mac Whinney (1982) for further discussion on developmental or staged learning.

[2] Such methods have also found application on topic spotting (Rose *et al.*, 1991), email sorting (Geutner *et al.*, 1993), and information retrieval (Van Rijsbergen, 1977).

[3] The reader is referred to Sutton (1992) and the background section of Gorin *et al.* (1991) for a discussion of the reinforcement learning literature.

[4] We refer the reader to Harnad (1990) and the background section of Sankar and Gorin (1993) for a discussion of the literature in language grounding.

[5] Distributional distances for within-language analysis are discussed, for example in Pereira and Tishby (1992).

[6] The reader is referred to the background section in Gorin *et al.* (1994a) for a discussion of this literature.

[7] There is a vast literature on this subject (Angluin and Smith, 1983) (Fu and Booth, 1975) (Jelinek, 1990) (MacWhinney, 1987) (Miclet, 1990) (Vidal *et al.*, 1994), and we refer the reader to the background section of Gorin *et al.* (1991) for further discussion.

[8] There has been much research into human language acquisition, which is beyond the scope of this paper. We refer the reader to Locke (1993) and Wanner and Gleitman (1982) for further discussion.

Aho, A. V., and Ullman, J. D. (1972). *The Theory of Parsing, Translation and Compiling, Vol. 1: Parsing* (Prentice-Hall, Englewood Cliffs, NJ).

Angluin, D., and Smith, C. H. (1983). "Inductive Inference: Theory and Methods," Comput. Surveys 15(3), 237–269.

Belkin, N. J., and Croft, W. B. (1992). "Information Filtering and Information Retrieval," Comm. ACM 35(12), 29–38.

Blachman, N. M. (1968). "The Amount of Information that y gives about X," IEEE Trans. Inform. Theory 14(1), 27–31.

Bunge, M. (1986). "A Philosopher Looks at the Current Debate on Language Acquisition," in *From Models to Modules*, edited by I. Gopnik and M. Gopnik (Ablex, NJ), Chap. 13.

Chomsky, N. (1965). *Aspects of the Theory of Syntax* (MIT, Cambridge, MA).

Cover, T. M., and Thomas, J. A. (1991). *Elements of Information Theory* (Wiley, New York).

Duda, R. O., and Hart, P. E. (1973). *Pattern Classification and Scene Analysis* (Wiley–Interscience, New York).

Elman, J. L. (1991). "Incremental Learning, or The Importance of Starting Small," Tech. Rep. 9101, Center for Research in Language, University of California at San Diego.

Farrell, K., Mammone, R., and Gorin, A. L. (1993). "Adaptive Language Acquisition Using Incremental Learning," Proc. of ICASSP, 1, 501–504 (April).

Fischer, K. W. (1980). "A Theory of Cognitive Development: The Control and Construction of Hierarchies of Skills," Psychol. Rev. 87, 477–525.

Flanagan, J. L., Berkley, D. A., and Shipley, K. L. (1991). "A Digital Teleconferencing System with Integrated Modalities for Human/Machine Communication: HuMaNet," Proc. ICASSP '91, Toronto, Canada 5, 3577–3579.

Fu, K. S., and Booth, T. L. (1975). "Grammatical Inference: Introduction and Survey—Parts I and II," IEEE Trans. on Systems, Man and Cybernetics SMCS, 95–111 and 409–423.

Gertner, A. N., and Gorin, A. L. (1993). "Adaptive Language Acquisition for an Airline Information Subsystem," *Artificial Neural Networks for Speech and Vision*, edited by R. Mammone (Chapman and Hall, London), pp. 401–428.

Geutner, P., Bodenhausen, U., and Waibel, A. (1993). "Flexibility Through Incremental Learning: Neural Networks for Text Categorization," in *Proceedings of the World Congress on Neural Networks*, Portland, Oregon.

Good, I. J. (1953). "The Population Frequencies of Species and the Estimation of Population Parameters," Biometrika 40, 237–264.

Goodman, R. M., Higgins, C. M., Miller, J. W., and Smyth, P. (1992). "Rule-Based Neural Networks for Classification and Probability Estimation," Neural Comput. 4, 781–804.

Gorin, A. L., and Levinson, S.E. (1989). "A Neural Network with Information-Theoretic Connection Weights," AT&T Bell Laboratories Technical Memorandum.

Gorin, A. L., Levinson, S. E., Gertner, A., and Goldman, E. (1991). "Adaptive Acquisition of Language," Comput. Speech Lang. 5(2), 101–132.

Gorin, A. L., Levinson, S. E., and Miller, L. G. (1993a). "Some Experiments in Spoken Language Acquisition," Proc. of ICASSP 1, 505–509.

Gorin, A. L., Wilpon, J. G., Sankar, A., and Miller, L. (1993b). "Spoken Language Acquisition for Automated Call Routing in a Telephone Network," in *Proceedings of the Workshop on Automated Speech Recognition* (Snowbird, Salt Lake City).

Gorin, A. L., Levinson, S. E., and Sankar, A. (1994a). "An Experiment in Spoken Language Acquisition," IEEE Trans. Speech Audio 2(1), Pt. II, 224–240.

Gorin, A. L., Hanek, H., Rose, R., and Miller, L. (1994b). "Spoken Language Acquisition for Automated Call Routing," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Yokohama (Acoustical Society of Japan, Tokyo), pp. 1483–1486.

Hanek, H. (1994). Private communication.

Harnad, S. (1990). "The Symbol Grounding Problem," Physica D 42, 335–346.

Henis, E. A., Levinson, S. E., and Gorin, A. L. (1994). "Mapping Natural Language and Sensory Information into Manipulatory Actions," in *Proceedings of the Yale Workshop on Adaptive and Learning Systems* (Yale University, New Haven).

Itakura, F. (1975). "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust. Speech, Signal Process. ASSP-23, 67–72.

Jelinek, F. (1990). "Self-Organizing Language Modeling for Speech Recognition," in *Readings in Speech Recognition*," edited by A. Waibel and K. Lee (Morgan-Kaufmann), pp. 450–506.

Kuhl, P. K. (1992). "Infants' Perception and Representation of Speech: Development of a New Theory," in *Proceedings of the International Conferenence on Spoken Language Processing (ICSLP)*, Alberta, Canada (ARPA, Washington, DC), pp. 449–456.

Lee, S. (1993). Private communication.

Levenshtein, V. I. (1966). "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," Cybernet. Control Theory 10(8), 707–710.

Levinson, S. E. (1985). "Structural Methods in Automatic Speech Recognition," Proc. IEEE 73, 1625–1650.

Levinson, S. E., and Shipley, K. L. (1980). "A Conversational Mode Airline Information and Reservation System Using Speech Input and Output," Bell Syst. Tech. J. 59, 119–137.

Locke, J. L. (1993). *The Child's Path to Spoken Language* (Harvard U.P., Cambridge, MA).

MacWhinney, B. (1982). "Basic Syntactic Processes," in *Language Acquisition: Vol I. Syntax and Semantics*, edited by S. Kuczaj (Erlbaum, Hillsdale, NJ).

MacWhinney, B. (Ed.) (1987). *Mechanisms of Language Acquisition* (Erlbaum, Hillsdale, NJ).

Maratsos, M. (1982). "The Child's Construction of Grammatical Categories," in *Language Acquisition: The State of the Art*, edited by E. Wanner and L. R. Gleitman (Cambridge U. P., Cambridge), pp. 240–266.

Masukata, M., and Nakagawa, S. (1994). "Concept and Grammar Acquisition Based on Combining with Visual and Auditory Information," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Yokohama (ASJ, Tokyo), pp. 1165–1166.

McDonough, J., *et al.* (1994). "Approaches to Topic Identification on the Switchboard Corpus," to appear in Proc. ICASSP '94.

McDonough, J., and Gish, H. (1994). "Issues in Topic Identification on the Switchboard Corpus," *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Yokohama (ASJ, Tokyo), pp. 2163–2166.

Miclet, L. (1990). "Grammatical Inference," in *Syntactic and Structural*

Pattern Recognition and Applications, edited by H. Bunke and A. Sanfeliu (World Scientific, Singapore).

Miller, L. G., and Gorin, A. L. (1993a). "A Conversational-Mode Automated Call Routing System," AT&T Bell Laboratories Tech. Memo.

Miller, L. G., and Gorin, A. L. (1993b). Spoken Language Acquisition in an Almanac Data Retrieval Task," AT&T Bell Laboratories Tech. Memo.

Miller, L. G., and Gorin, A. L. (1993c). "Structured Networks for Adaptive Language Acquisition," Intl. J. Pattern Recognition Artificial Intelligence 7(4), 873–898.

Minsky, M. L., and Papert, S. A. (1990). Perceptrons (MIT, Cambridge, MA).

Pereira, F. (1994). Private communication.

Pereira, F., and Tishby, N. (1992). "Distribution Similarity, Phase Transitions and Hierarchical Clustering," in *Proceedings of the AAAI Symposium on Prob. Approach to Natural Language Processing II*, Cambridge, pp. 108–112.

Peskin, B. (1993). "Topic and Speaker Identification via Large Vocabulary Speech Recognition," in *Proceedings of the ARPA Workshop on Human Language Technology* (ARPA, Washington, DC).

Pieraccini, R., *et al.* (1992). "Progress Report on the Chronus System: ATIS Benchmark, in *Proceedings of the DARPA Workshop on Speech and Natural Language* (ARPA, Washington, DC), Session 3.

Pierce, J. R. (1961). *Symbols, Signals and Noise* (Harper, New York).

Prieto, N., and Vidal, E. (1992). "Learning Language Models through the ECGI Method," Speech Commun., 299–309.

Rabiner, L. R., and Juang, B. H. (1993). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).

Rabiner, L. R., Wilpon, J. G., and Soong, F. K. (1989). "High Performance Connected Digit Recognition using HMMs," IEEE Trans. ASSP Acoust. Speech Signal Process. ASSP-37(8), 1214–1225.

Richard, M. D., and Lippmann, R. (1991). "Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities," Neural Comput. 3, 461–483.

Rohlicek, J. R., *et al.* (1992). "Gisting Conversational Speech," Proc. ICASSP '92, II, 113–116.

Rose, R., Chang, E. I., and Lippmann, R. (1991). "Techniques for Information Retrieval from Speech Messages," Proc. ICASSP, Vol. 1, 317–320.

Rose, R. C. (1993). "Definition of Acoustic Subword Units for Word Spotting," in *European Conference on Speech Communication and Technology*.

Rulot, H., and Vidal, E. (1987). "Modelling (Sub)string-Length-Based Constraints through a Grammatical Inference Method," in *Pattern Recognition: Theory and Applications*, edited by P. Devijver and J. Kittler (Springer-Verlag, Berlin), pp. 451–459.

Rumelhart, D.E., and McClelland, J. L. (Eds.) (1988). *Parallel Distributed Processing* (MIT, Cambridge, MA), Vol. 1.

Sankar, A., and Gorin, A. L. (1993). "Adaptive Language Acquisition in a Multisensory Device," in *Artificial Neural Networks for Speech and Vision*, edited by R. Mammone (Chapman and Hall, London), pp. 324–356.

Sankar, A., Gorin, A. L., Wilpon, J. G., Lee, S. Y., Venkataramani, R., and Bock, D. (1993). "Language Acquisition for Automated Call Routing in the Telephone Network," AT&T Bell Laboratories Tech. Memo.

Sankoff, D., and Kruskal, J. B. (1983). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* (Addison-Wesley, Reading, MA).

Shannon, C. E. (1948). "A Mathematical Theory of Communication," Bell Syst. Tech. J. 27(3), 379–423.

Shannon, C. E. (1951). "Prediction and Entropy of Printed English," Bell Syst. Tech. J., 50–64 (Jan.).

Sutton, R. (1992). "Introduction: The Challenge of Reinforcement Learning," *Machine Learning, Special Issue on Reinforcement Learning* 8, 225–227.

Tishby, N. Z., and Gorin, A. L. (1994). "Algebraic Learning of Statistical Associations," Comput. Speech Lang. 8(1), 51–78.

Van Rijsbergen, C. J. (1977). "A Theoretical Basis for the use of Co-occurrence Data in Information Retrieval," J. Documentation 33(2), 106–119.

Vidal, E. (1992). Private communication.

Vidal, E., Casacuberta, F., and Garcia, P. (1994). "Syntactic Learning Techniques for Language Modeling and Acoustic Phonetic Decoding," in *Speech Recognition and Coding. New Advances and Trends*, edited by J. Rubio and J. M. Lopez (Springer-Verlag, Berlin).

Wanner, E., and Gleitman, L. R. (1982). *Language Acquisition: The State of the Art* (Cambridge U.P., Cambridge).

Ward, W. (1991). "Understanding Spontaneous Speech," Proc. ICASSP '91, 365–357.

Watt, W. C. (1968). "Habitability," Am. Documentation, 338–351 (July).

Webster, D., (1987). *Webster's New Collegiate Dictionary* (Merriam-Webster, Springfield, MA).

Wilpon, J., Rabiner, L. R., and Martin, T. (1984). "An Improved Word Detection Algorithm for Telephone Quality Speech Incorporating Both Syntactic and Semantic Constraints," AT&T Tech. J. 63(3), 479–498.

Winograd, T. (1983). *Language as a Cognitive Process, Volume I: Syntax* (Addison-Wesley, Reading, MA).

Woods, W. A. (1968). "Procedural Semantics for a Question Answering Machine," *Proceedings of AFIPS*, pp. 457–471.

Yarowsky, D. (1990). Private communication.

Zipf, G. K. (1949). *The Principle of Least Effort* (Addison-Wesley, Reading, MA).

Zue, V. (1992). "The MIT ATIS System," in *Proceedings of the DARPA Workshop on Speech and Natural Language* (ARPA, Washington, DC), Session 3.